A Power Law for Reduced Precision at Small Spatial Scales: Experiments with an SQG Model

Tobias Thornes*, Peter Düben† & Tim Palmer*†

*Atmospheric, Oceanic and Planetary Physics, University of Oxford, UK

†European Centre for Medium-Range Weather Forecasts, UK

Contact: tobias.thornes@oriel.ox.ac.uk

Abstract

Representing all variables in double-precision in weather and climate models may be a waste of computer resources, especially when simulating the smallest spatial scales, which are more difficult to accurately observe and model than are larger scales. Recent experiments have shown that reducing to single-precision would allow real-world models to run considerably faster without incurring significant errors. Here, the effects of reducing precision to even lower levels are investigated in the Surface Quasi-Geostrophic system, an idealised system that exhibits a similar power-law spectrum to that of energy in the real atmosphere, by emulating reduced precision on conventional hardware. It is found that precision can be reduced much further for the smallest scales than the largest scales without inducing significant macroscopic error, according to a -4/3 power law, motivating the construction of a 'scale-selective' reduced-precision model that performs as well as a double-precision control in short- and long-range forecasts but for a much lower estimated computational cost. A similar scale-selective approach in real-world models could save resources that could be re-invested to allow these models to be run at greater resolution, complexity or ensemble size, potentially leading to more efficient, more accurate forecasts.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.3303

1. Introduction

More accurate weather and climate forecasts could be achieved by increasing the resolution, complexity and ensemble size of the operational numerical models that underlie them. But these factors are limited by the size of the supercomputers that forecast centres can afford. Because the density of transistors in conventional chips can be increased no further without the chips overheating, individual computer processors are no longer getting faster (Markov 2014), so that more powerful supercomputers are usually constructed by running more processors in parallel, increasing the energy running-costs proportionally. Resolving individual convective clouds, which are a key source of forecast uncertainty, would require a fifty-fold increase in the effective resolution of today's best models, but the computers required for this may remain unaffordable for many years (Palmer 2014).

For this reason, ways are being sought to increase the efficiency of hardware, to provide better forecasts at today's energy costs. One means is to use 'mixed precision' inexact hardware capable of representing different computations with different levels of numerical precision. Conventionally, all calculations are done in 'double-precision'. According to Institute of Electrical and Electronics Engineers standards (IEEE 2008) a double-precision number is represented during computations using 64 bits of information: 1 bit for the sign, 11 bits for the 'exponent' (the nearest power of 2) and 52 bits for the 'significand' (a fraction between 1 and 2), where the overall number is formed by multiplying these components together. Numbers of up to 15 decimal digits are represented exactly, so long as they fall between 2^{-1022} (or lower if exponent bits are borrowed from the significand to represent 'sub-normal' numbers) and 2^{1023}.

Reducing the number of exponent bits relative to double-precision reduces the range of representable numbers, whilst reducing the number of significand bits reduces the numerical resolution. In 'single-precision', for example, there are 32 bits in total: only 8 bits are given to the exponent and 23 bits to the significand, which means that only numbers between 2^{-126} and 2^{127} can be represented, with up to 7 decimal places. 'Half-precision' uses 16 bits in total, with just 5 bits for the exponent and 10 bits for the significand. However, reducing precision also reduces the time taken

to carry out computations, move information between processors and store data, which can lead to more efficient computations so long as numbers can still be represented sufficiently accurately.

Recently, it was shown that running the European Centre for Medium-Range Weather Forecasts (ECMWF)'s operational model, the Integrated Forecasting System (IFS), almost entirely in single-precision yielded 40% savings in computational cost relative to double-precision without impeding the accuracy (Vana et al. 2017). Further savings could be achieved by running at even lower precision, but this could incur considerable rounding errors if done indiscriminately. Here, we explore the possibility of avoiding such errors by taking a scale-selective approach, reducing precision more for calculations that pertain to smaller spatial scales in spectral models. At these scales, numerical precision may be less critical because the observations used to initialise models are of limited spatial resolution and so inherently imprecise at small scales; and a model's output is less accurate close to the truncation scale, which is why relatively large random perturbations are applied to these scales when stochastic parameterisation schemes are used in ensemble forecasts. This is consistent with the theoretical prediction that in a chaotic atmosphere errors at smaller scales are less critical to forecasts than those at larger scales because they are quickly swamped by down-scale error propagation (Durran and Gingrich 2014).

Indeed, in a study of an idealised system based on Lorenz' 1996 model (2006), we demonstrated that reducing the smallest resolved scales to half-precision and larger scales to single-precision yielded a more accurate forecast than reducing the model resolution, for similar computational costs (Thornes et al. 2017). Düben, McNamara and Palmer found that scale-selective precision could also be beneficial for a spectral dynamical core (Düben et al. 2014).

In this study, we build upon such earlier work to investigate the effects of scale-selective reduced precision in greater detail, using an idealised atmosphere based on the Surface Quasi-Geostrophic (SQG) equations. The results could be used to motivate and inform future studies on scale-selective precision in real-world spectral models such as IFS, where reducing precision to its optimal level at each spatial scale would avoid wasting computational resources that could instead be used to increase the model resolution or complexity. In the absence of mixed-precision hardware capable of reducing to arbitrary precision, we emulated reduced precision on conventional hardware,

making it difficult to estimate the actual cost savings. Nonetheless, processing units capable of running in double-, single- and half-precision have already been developed (Trader 2016), and experiments on Field Programmable Gate Arrays, which have specifiable precision but are currently too complex to program to run full atmospheric models in mixed-precision, suggest that reducing to even less than half-precision would yield further efficiency improvements (Russell et al. 2015).

In Section 2 of this paper, we describe the SQG model and its relevance to the real atmosphere. Section 3 introduces the program that we used to emulate reduced-precision for our experiments in the SQG system, the results of which are described in Section 4. Section 5 describes the maximum potential cost savings, and the implications of these results are discussed in Section 6.

2. The Surface Quasi-Geostrophic Model

The Surface Quasi-Geostrophic (SQG) Model is based on a simplification of the Navier-Stokes equations that can be used to describe the evolution of the real-Earth atmosphere. Quasi-Geostrophic (QG) theory is applicable to situations of approximate geostrophic and hydrostatic balance between pressure-gradient, Coriolis and gravitational forces, and describes the dynamics of small deviations from this balance in a three-dimensional rapidly rotating atmosphere. The flow can be described by the two-dimensional wind components u and v, described using the streamfunction ψ, where,

$$(u, v) = \left(-\frac{\partial \psi}{\partial y}, \frac{\partial \psi}{\partial x}\right). \tag{1}$$

The buoyancy or potential temperature, Θ, is defined as,

$$\Theta = \frac{\partial \psi}{\partial z}, \tag{2}$$

and the vorticity ζ, is,

$$\zeta = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2}. \tag{3}$$

Any of these key variables can be used to drive the model and evaluate its output.

SQG is derived from QG by further assuming that the potential vorticity q is constant (Held et al. 1995). This quantity is a three-dimensional extension of the vorticity, and is defined as,

$$q = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial}{\partial z}\left(\epsilon \frac{\partial \psi}{\partial z}\right), \tag{4}$$

where $\epsilon = (f/N)^2$. Here, N is the buoyancy frequency and f is the Coriolis parameter $f = 2\Omega \sin(\phi)$, which describes the background rotational vorticity of the fluid at latitude ϕ and rotation rate Ω. Hence, a constant $q = q_0$ means that the variation in ψ can be predicted entirely from its behaviour in x and y at the bottom surface of the atmosphere, and the solution is assumed to decay exponentially with z above this surface (Tulloch and Smith 2006). The SQG equations therefore apply strictly to a two-dimensional plane – the 'surface' – but are quasi-three-dimensional in that they encapsulate the behaviour of a three-dimensional fluid.

In a Boussinesq (constant density) SQG fluid, the vorticity, buoyancy and streamfunction on the bottom surface evolve according to,

$$\frac{\partial \zeta}{\partial t} = \frac{\partial \zeta}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial \psi}{\partial x}\frac{\partial \zeta}{\partial y}, \tag{5}$$

$$\frac{\partial \Theta}{\partial t} = \frac{\partial \Theta}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial \psi}{\partial x}\frac{\partial \Theta}{\partial y}, \tag{6}$$

$$\psi = -\int \frac{1}{|y|}\Theta(x+y,t)dy. \tag{7}$$

If the bottom surface is not flat, but incorporates some topographical function $h(x, y)$, the buoyancy is modified so that (Held et al. 1995),

$$\frac{\partial \Theta}{\partial t} = \frac{\partial \psi}{\partial y}\frac{\partial}{\partial x}(\Theta + N^2 h) - \frac{\partial \psi}{\partial x}\frac{\partial}{\partial y}(\Theta + N^2 h). \tag{8}$$

In this investigation, the SQG equations were solved numerically using a Fortran computer program based on code developed by Smith (2009). Equation (8) is discretised in time, using adaptive time-stepping and the Runge-Kutta scheme on a periodic domain. The time-dependent left-hand side is solved in spectral space, where different scales are represented through different wavenumbers k. But solving the non-linear terms in the right-hand-side in spectral space would require cross-multiplying between all wavenumbers, so it is much more efficient to solve this part in grid-point space (Bourke 1972). The program uses Fast Fourier Transforms (FFTs) to change between these configurations, using the well-known 'FFTW' library.

Our SQG system uses a two-dimensional spectral plane, so that every spectral variable is described by two wavenumber components, in the zonal (k_x) and meridional (k_y) directions. The total wavenumber for each basis function is therefore given by $k = \sqrt{k_x^2 + k_y^2}$,. We used a $2\pi \times 2\pi$

periodic domain, with a spectral resolution $k_{\max} = 127$, meaning that there are 128 wavenumbers (including 0) in each of the k_x- and k_y- directions. For quantities evaluated in grid-point space, this corresponds to a resolution of $2k_{\max}$ grid-points in each of the x- and y- directions. In the real Earth atmosphere, energy is added mostly at large scales and removed at small scales by viscous dissipation. To simulate the latter, we switched on eighth-order hyperviscosity in the existing code. This removes energy by damping down the smallest scales and varies according to,

$$\Theta(k) = \Theta(k)\left[1 + d\left(\frac{4\pi k}{k_{\max}}\right)^8\right]^{-1}, \tag{9}$$

where d is the damping strength.

The single-level nature of our SQG model means that the Coriolis parameter f is effectively set to zero in the model code. Hence, to create Earth-like external forcing we introduced a sinusoidally varying zonal background wind $u_b = u_0 \sin(\phi)$, which falls to zero at the edges of the domain in the zonal direction and is constant along the meridional direction, alongside realistic, fractal topography of mean height h_0. For investigations of short-term forecasts, the setup included a large central protrusion resembling a mountain of height H rising above the rough terrain. This was made high enough to interrupt the flow of the simulated atmosphere without blocking it entirely, simulating real-world flow over a mountain. All the variables in the model are nondimensionalised, so that the two-dimensional position and velocity vectors of a fluid parcel in the model are $\hat{x} = \vec{x}/L$ and $\hat{u} = \vec{u}/U$ respectively, where L and U are the dimensional length and speed that can be specified to translate the model variables into real quantities. The essentially infinite vertical scale in our experiments means that there is no Rossby number with which to relate the system to the real world, but preliminary experiments revealed that setting the parameters to $d = 10$, $u_0 = 0.5$, $h_0 = 0.5$ and $H = 3$ produced weather-like features. An example snapshot of the vorticity ζ field for this setup is shown in Figure 1, where the central depression corresponds to the mountain, around which the fluid is flowing. The forcing from the mountain means that the total energy in this setup continues to grow indefinitely; for longer-term forecast investigations where a stable equilibrium is desirable we therefore used a setup that had no mountain and had $u_0 = 0.02$ to conserve total energy, but was otherwise identical.

SQG is especially interesting for a scale-selective precision study because it exhibits wavenumber power spectra for the energy per unit wavenumber E and enstrophy per unit wavenumber G similar to those observed in the real atmosphere, where it has been observed that $E \propto k^{-5/3}$ on small scales – on the order of 100km (Nastrom and Gage 1985), depending on height – and $E \propto k^{-3}$ on larger scales. These regimes have been referred to as '3D' and '2D' turbulence by some (Leith 1971) and correspond to energy dissipating down to small scales and up to larger scales respectively. In SQG, there is a downscale cascade for the enstrophy G that follows $G \propto k^{-5/3}$, whilst $E \propto k^{-8/3}$, a slightly steeper downscale energy cascade, at high wavenumbers.. Theory predicts an upscale energy cascade at low wavenumbers (large scales), where $E \propto k^{-2}$ (Pierrehumbert et al. 1994).

Figure 2 shows the spectra we measured for the energy and enstrophy per unit wavenumber (E and G) with our short-term forecasting setup. In both cases, only the downscale energy cascade is resolved, but the spectra exhibit approximately the expected slopes of -8/3 and -5/3 respectively. The similar downscale cascade between SQG and the real atmosphere may mean that the propagation of observational and modelling errors is also similar, something that we hope to explore in more detail in a forthcoming paper. By measuring the rate of error growth at wavenumbers corresponding roughly to the synoptic forecasting scale in the real atmosphere ($k \approx 10$) to the time taken for initial errors to double in size (the error doubling time) of typical real-Earth models, which Lorenz (2006) estimated to be around 1.5 days (in 1996), we were able to estimate that one model timestep in our setups (at $k_{\max} = 127$ resolution) is roughly equivalent to 2 minutes in a full-complexity atmospheric model.

3. Emulating Reduced Precision

Because real hardware capable of scale-selectively reducing precision to less than 16 bits of precision in an atmospheric model is not yet widely available, it was necessary for us to emulate reduced precision on double-precision processors by using an emulator program developed by our research group (Dawson and Düben 2016). This emulator rounds numbers in the calculations to which it is applied as though they were represented with a specified number of bits. Values that fall outside the

range representable with a reduced-precision exponent are rounded to zero or infinity. The program allows the number of bits to be specified for each wavenumber component of each variable individually, allowing for mixed-precision across different spatial scales.

Since there is no IEEE standard for the ratio of significand to exponent bits below half-precision, we used the single-precision exponent (8 bits) throughout our experiments for convenience. Running with the half-precision exponent (5 bits) instead in reduced-precision runs made no difference to the results, and we deduce from this that the loss of resolvable range with only 5 bits is not problematic for the SQG model.

During calculations, our main SQG program calls on the emulator to round numbers that have been assigned a special reduced-precision Fortran 'type' as though they were in reduced precision before and after every operation. Doing this means that, rather than speeding up calculations as real reduced-precision hardware would do, using an emulator slows them down. It is therefore not possible for us to measure the actual computational cost saving achievable with reduced precision. However, assuming that the cost of a calculation would be proportional to the number of bits used yields an upper estimate on the savings achievable, which might be realised by the most efficient imaginable hardware. In reality, there would be additional costs, for example those associated with specifying what level of precision to use, lowering the potential savings.

4. Reduced Precision Experiments

We carried out a series of experiments to determine whether scale-selectively reducing precision could reduce computational cost without increasing the model error in both short-term and long-term forecasts of the SQG system. At the beginning of each experiment, the model was spun-up for 50000 model timesteps from zeroed initial fields using the features described in Section 2 for our setups. The state of the variables at 50000 timesteps defined new 'initial conditions'. From this point, the 'truth' system was run forward and its time-evolution was compared to that of each of a set of 'models' forecasting the system at different levels of precision. The observational uncertainty inherent in real-world models was simulated by running an ensemble of independent forecasts for each model, each of

which began with the 'truth' initial conditions altered slightly by randomly perturbing each of the streamfunction grid-points by up to $\pm 1\%$.

Three experiments were performed, as outlined below. For short-term forecasts, the metric for comparing the model to the 'truth' we used was the 'Absolute Error', computed at each output timestep by measuring the absolute difference between the 'truth' and model values of the streamfunction or vorticity at each point in grid-point space and averaging over all grid-points. This measure is simple to compute but provides a clear indication of the onset of differences between the reduced-precision and double-precision models. Once the lowest precision levels that could be used accurately had been identified, further metrics such as the Root Mean Square Error (RMSE) and the change to the ensemble spread were used to verify the performance of the resulting optimised-precision system, as described in Section 4.2 below.

For longer-term forecasts, beyond the maximum lead-time at which the forecast is still correlated to the true state of the system, the Kolmogorov-Smirnov (KS) and Hellinger Distance statistics were used to measure the difference in mean conditions over the course of a long integration. For these, the measured variable – in our case, the streamfunction – at every output timestep for each ensemble member is allocated to one of a set of 100 bins according to its magnitude; the overall populations of the bins then represent a probability distribution function (pdf) that describes how often the system occupies each of the 100 streamfunction states. The KS statistic assesses a model's performance by measuring the maximum difference between the pdfs of the model and the 'truth' out of all the bins, whereas the Hellinger Distance integrates the difference across all bins (Pollard 2002).

4.1. Optimising Spectral Precision

The first experiment aimed to determine how low precision could be reduced at each spatial scale in a short-term forecast of the system for there to be negligible error in the overall streamfunction or vorticity. The system was run for 5000 timesteps after spin-up, which corresponds to around 7 'days'. This 'truth' was compared against a control ensemble of fifty double-precision forecasts, produced as outlined above, to yield fifty Absolute Error timeseries. The ensemble mean was taken as the Control Error, and reflects the model error that would be expected in the absence of reduced precision

rounding errors. The ensemble standard deviation was used to estimate the uncertainty in this mean (Hughes and Hase 2010).

To investigate the additional error introduced by reducing to a given number of significand bits s, the ensemble mean Absolute Error was also computed for a fifty-member ensemble of reduced-precision forecasts for which all wavenumbers below the truncation wavenumber k_t were in single-precision and all wavenumbers above k_t were reduced to s bits in the significand. Several different values of s were tried between 1 and 23, and for each of these k_t was varied. The larger the value of k_t, the smaller the scales to which reduced precision is confined, and the less potential there is for rounding error. For each s, we determined the smallest value of k_t for which the reduced precision and control model mean forecasts agreed to within the experimental uncertainty of the control after 3000 timesteps (4 'days'). Table 1 shows the results. Although in all cases precision was reduced to the same level for all spectral variables, the Absolute Error was measured separately for the streamfunction ψ and vorticity ζ. The minimum values of k_t were different in these two cases because the vorticity is a more local property that exhibits smaller-scale features than the streamfunction, and so is more sensitive to reducing precision at high wavenumbers.

The results accord with the hypothesis that smaller spatial scales require less precision for the overall forecast to be accurate. Whilst degrading the largest scales ($k_t < 2$) to less than 18 bits leads to an inaccurate streamfunction forecast, wavenumbers above $k_t = 111$ for ψ or $k_t = 123$ for ζ can be represented with just 5 bits in the significand with negligible impact. The smallest scales are nonetheless necessary, as reducing even the highest wavenumber to less than 5 significand bits or omitting it entirely was sufficient to produce significant errors, which implies that lowering the resolution of the model to save costs would not be as accurate as is scale-selectively reducing precision. The precision (number of significand bits) required, s, in the streamfunction and vorticity separately is plotted logarithmically against wavenumber in Figure 3. This yields a straight line, the best-fit for which in the streamfunction has a gradient $m = -0.30 + 0.01$ and intercept $c = 3.12 \pm 0.03$, and in the vorticity has $m = -0.27 \pm 0.01$ and $c = 3.01 \pm 0.04$computed using the GNUPLOT computer program, which implies that, to a close approximation, $s \propto k^{-1/3}$.

4.2 Scale-Selective Models

In light of these results, our second experiment aimed to determine whether a scale-selective model could yield accurate short-term forecasts. We therefore performed longer, twenty-five member ensemble forecasts that compared the control model error to that of a scale-selective model for which precision was reduced for each range of wavenumbers to the level required for negligible Absolute Error according to Table 1. The forecasts were run for five sets of initial conditions and the results were averaged across these. Models with uniformly reduced precision were also run for comparison. Figure 4 shows the Root Mean Square Error (RMSE) in each of ψ and ζ when spectral precision was optimised to minimise error in ζ, which is generally the more sensitive of the two parameters. The RMSE is shown in preference to the Absolute Error because it is in general more sensitive to errors at small spatial scales, but in this case the two measures yielded identical results.

The scale-selective runs are indistinguishable from the double-precision control for up to 35000 timesteps in the streamfunction and 20000 timesteps in the vorticity, implying that scale-selective precision preserves accuracy in forecasts for lead-times of days or weeks of model time. A uniform reduction in precision even to 15 bits, on the other hand, produces significant error throughout. Figure 5 compares the vorticity fields visually at 30000 timesteps after spin-up, demonstrating that most features are captured with very little change relative to the double-precision control when precision is reduced, but that the 15 bit case matches the double-precision slightly less accurately than does the scale-selective model. Further tests showed that reducing the number of significand bits by 1 at each spatial scale relative to the scale-selective model degraded the accuracy slightly but consistently, and that this additional error was similar even when only smaller scales ($k > 15$) were reduced in precision relative to the scale-selective model. This implies that our scale-selective model indeed uses the optimal number of bits at each spatial scale for our system setup.

The same three models were also compared for longer-term forecasts of the mean state of the model's 'climate'. To do this, the system setup with no mountain was run for 50000 timesteps, then initial conditions were randomly perturbed as in the previous experiment and the same three models were run forward and compared against the 'truth' for 250000 timesteps (around 1 'year' of model

time), beginning after another 250000 timesteps had elapsed to allow the models to stabilise, using the KS and Hellinger Distance statistics. Each model was run three times with different random perturbations to compute its pdf, which was averaged over the three ensemble members in each case before being compared to the single 'truth' pdf. KS and Hellinger statistics were computed in this way for four independent runs differing in the (random) bottom topography and hence the initial conditions, and the mean and standard error were calculated across the four runs.

The results are shown in Figure 6. The scale-selective and double-precision models have the same long-term forecasting accuracy to within the experimental error, whilst the 10 bit model performs very poorly by comparison. Further tests using single runs (to save computational time) of other reduced-precision models suggested that at least 15 bits in the significand would be required for negligible error when applying uniform reduced precision across all spatial scales.

4.3 Grid-point and FFT Components

It is not possible to apply scale-selective precision during grid-point calculations, where values are not computed separately for different spatial scales, or in the Fast Fourier Transform (FFT), where different scales are mixed during calculations. From Table 1, we can assume that for grid-point calculations, the lowest level of precision that can be used uniformly without error is 19 bits.

To compute the acceptable precision level in the FFTs, we wrote a new FFT algorithm based on that provided in Press et al. (1992), to which we introduced the reduced-precision Fortran 'type' required to apply our emulator: the much faster FFTW algorithm usually used by the model is not compatible with the reduced-precision type. When the streamfunction Absolute Error was computed for short-term (2000 timestep) runs of a model with reduced precision in the FFT part only, it was found that reducing the significand to 11 bits produced no noticeable increase in streamfunction error relative to the double-precision control, whereas reducing to 10 bits in the significand (for the same single-precision exponent) caused considerable errors, implying that the algorithm is not affected by small rounding errors, but crashes entirely when such errors are sufficiently large.

Furthermore, we found that a reduction to 11bits and 19bits for the FFTs and grid-point calculations respectively could be combined with scale-selective spectral-space precision without

significantly increasing the streamfunction or vorticity Absolute Error or RMSE relative to representing the FFTs and grid-point calculations in double-precision. The streamfunction RMSE result is shown in Figure 7 for a ten-member ensemble with a single initial condition; the Absolute Error and the corresponding vorticity plots give almost identical results and are omitted for conciseness. The model remains accurate for up to 30000 timesteps, which is of the order of 40 atmospheric days, well beyond the range of an accurate weather forecast. To verify that reduced precision in one or more of the components does not significantly increase the ensemble spread, we also tested the impact on ensemble spread of the scale-selective models with and without reduced-precision in the FFT and grid-point components, with streamfunction results also shown in Figure 7. These models do not exhibit a sustained increase in ensemble spread relative to the double-precision control, and there is no difference in spread at all between any of the models until 40000 timesteps have elapsed.

5. Discussion

The results presented in Section 4 imply that precision can be reduced to 19 bits in the grid-point part of the SQG system and 11 bits in the FFT component without inducing considerable error or increasing the ensemble spread, and that in the spectral component lower precision can be tolerated at smaller spatial scales. The physical explanation for the -1/3 power law with wavenumber of the required precision is unclear, but it may be related to the -5/3 power law of the enstrophy spectrum, which also describes the relative magnitude of the streamfunction as a function of wavenumber. This dropping off at higher wavenumbers means that smaller scales should make a smaller relative contribution to the overall streamfunction value when the Absolute Error is computed, which would imply that precision can be safely reduced to a greater degree on smaller scales.

On the other hand, other experiments that we conducted showed that the error doubling time also follows a $-5/3$ power law as a function of wavenumber; this implies that errors grow more rapidly at smaller scales, so initial errors in small scales might be expected to be more influential than those at larger scales, in which case the small scales might actually require higher precision. It may be that these two factors both apply but act contrary to one another, which we conjecture (without proof)

could be one explanation for the observed shallow negative power law $s \propto k^{-1/3}$ overall. The way in which errors spread between spatial scales in SQG, which may be relevant here, is itself a question of scientific interest that we plan to explore in a forthcoming paper. It may also be relevant that the magnitude of the rounding error e_r relative to double-precision induced when the number of bits in the significand is reduced to level p goes as $e_r \propto 2^{64-p}$, which means that the relationship between the error and the number of significand bits is not linear.

To evaluate the potential computational cost saving with the mixed-precision model, we multiplied together the cost c_i relative to double-precision of each level of precision i, and the number of 1D wavenumber components n_i that should be reduced to this level according to Table 1 as a fraction of the total number of 1D wavenumbers, $k_{\max}^2$. Where s_i is the number of significand bits in level i, this yields the fractional cost F_i of precision level i, given by,

$$F_i = \frac{(s_i+5+1)n_i}{64k_{\max}^2} = c_i \frac{n_i}{k_{\max}^2}. \quad (10)$$

Summing this over all i yields an overall cost estimate F relative to a fully double-precision simulation. It is assumed that a half-precision exponent of 5 bits is used at all spatial scales in the mixed-precision model (this is justified for at least the parts of the model with 10 significand bits or fewer), with 1 bit for the sign; and that the number of bits used to represent numbers in a calculation is directly proportional to that calculation's cost.

We obtained an overall fractional cost of $F = 0.2$ for both streamfunction and vorticity optimisation, which implies that scale-selective precision can provide an 80% cost saving for spectral calculations in SQG. Real reduced-precision hardware may have overhead costs associated with specifying which level of precision to use at which scale, or might not speed up in inverse proportion to the number of bits used, lowering the potential savings. Nevertheless, assuming that the most efficient conceivable scale-selective reduced-precision hardware was used, this would mean that spectral calculations, which in double-precision were responsible for 50% of the cost of the model, required just 20% of their former cost.

Table 2 indicates that the 35% of the model run-time associated with FFTs would require 27% of its former cost if run with 11 bits in the significand, and that the 15% of the model run in grid-

point space would require 39% its former cost if run with 19 significand bits. Overall, this implies a cost of $50\% \times 20\% + 35\% \times 27\% + 15\% \times 39\% = 25\%$ of the double-precision standard, a saving of 75%. Figure 8 shows this calculation diagrammatically.

6. Conclusions

Our experiments have shown that precision can be reduced much more strongly at smaller spatial scales without incurring significant error in the prognostic variables in SQG, with as few as 5 bits required in the smallest scales for accurate calculations. The SQG system is composed of spectral, grid-point and FFT components, and the gains achievable with scale-selective precision are only achievable in the spectral part, reducing the cost of this component by an estimated 80% relative to double-precision were idealised reduced-precision hardware to be used with no overhead costs. If precision is constrained to be the same for all scales, as is necessary for grid-point calculations, the saving is only 61%, whereas in the FFTs that switch between these components a saving of 73% is possible. This yields an overall potential cost saving of 76%.

Even if the actual savings achievable are less than our idealised estimates, scale-selective precision shows considerable advantage over a uniform reduction in precision: at least 19 bits are required at the largest scales in SQG, which, if applied uniformly, would be around twice as expensive as the scale-selective alternative; uniform single-precision would be three times as expensive (Vana et al. 2017). Although hardware capable of straightforwardly applying such a mixture of precision levels to a full atmospheric model is not yet available, our findings imply that its development and use for weather and climate forecasts based on spectral models could be worthwhile. With such hardware, the fact that spectral models represent less crucial, smaller-scale features separately to larger-scale ones may render them more efficient than grid-point alternatives. We aim to perform similar experiments to those presented here in the real-world ECMWF Open IFS model, which also has a spectral component in the dynamical core, to demonstrate the potential of scale-selective precision to improve the efficiency of real weather and climate forecasts at a time when producing more accurate forecasts is more necessary, but potentially also more expensive, than ever before.

Acknowledgements

The authors would like to thank Ray Pierrehumbert for advice concerning SQG and Andrew Dawson for his part in developing the emulator used in this investigation. Tobias Thornes is funded by a NERC grant (number NE/L002612/1); Tim Palmer and Peter Düben by an ERC grant (Towards the Prototype Probabilistic Earth-System Model for Climate Prediction, project reference 291406); and Peter Düben gratefully acknowledges funding from the ESIWACE project under grant number 675191. We would also like to thank two anonymous reviewers for their help in significantly improving this paper.

Supplementary Material

The 'namelist' values used to define the settings used for this investigation are provided as supplementary material in the file 'input.nml', which can be opened with any text editor.

References

Bourke W. 1972. An Efficient, One-Level, Primitive-Equation Spectral Model. *Mon. Weather Rev.* **100** : 683–689. doi:10.1175/1520-0493(1972)100<0683:AEOPSM>2.3.CO;2

Dawson A, Düben PD. 2016. rpe v5: An Emulator for Reduced Floating-point Precision in Large Numerical Simulations. *Geosci. Model Dev. Discuss.* doi:10.5194/gmd-2016-247

Düben PD, McNamara H, Palmer TN. 2014. The Use of Imprecise Processing to Improve Accuracy in Weather and Climate Prediction. *J. Comput. Phys.* **271** : 2–18. doi:10.1016/j.jcp.2013.10.042

Durran DR, Gingrich M. 2014. Atmospheric Predictability: Why Butterflies are Not of Practical Importance. *J. Atmos. Sci.* **71** : 3476–2488. doi:10.1175/JAS-D-14-0007.1

Held IM, Pierrehumbert RT, Garner ST, Swanson KL. 1995. Surface Quasi-Geostrophic Dynamics. *J. Fluid Mech.* **282** : 1–20. doi:10.1017/S0022112095000012

Hughes I, Hase T. 2010. *Measurements and their Uncertainties*. 1st ed. Cambridge University Press: Cambridge;

IEEE. 2008. IEEE Standard for Floating-Point Arithmetic. *IEEE Stand.* 1–70.

Leith CE. 1971. Atmospheric Predictability and Two-Dimensional Turbulence. *J. Atmos. Sci.* **28** : 145–161. doi:10.1175/1520-0469(1971)028<0145:APATDT>2.0.CO;2

Lorenz E. 2006. Predictability - a Problem Partly Solved. *Predictability of Weather and Climate*, R. Palmer T & Hagedorn, Ed., Cambridge University Press: eds. T. N. Palmer & R. Hagedorn, 40–58

Markov IL. 2014. Limits on Fundamental Limits to Computation. *Nature* **512** : 147–154.

Nastrom GD, Gage KS. 1985. A Climatology of Atmospheric Wavenumber Spectra of Wind and Temperature Observed by Commercial Aircraft. *J. Atmos. Sci.* **42** : 950–960.

Palmer TN. 2014. Build High-Resolution Global Climate Models. *Nature* **515** : 338–339.

Pierrehumbert RT, Held IM, Swanson KL. 1994. Spectra of Local and Non-Local Two Dimensional Turbulence. *Chaos, Solitons & Fractals* **4** : 1111–1116. doi:10.1016/0960-0779(94)90140-6

Pollard D. 2002. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press: Cambridge;

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1992. Fast Fourier Transform. *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*, Cambridge University Press: Cambridge, 490–529

Russell FP, Düben PD, Niu X, Luk W, Palmer TN. 2015. Architectures and precision analysis for modelling atmospheric variables with chaotic behaviour. *Proc. - 2015 IEEE 23rd Annu. Int. Symp. Field-Programmable Cust. Comput. Mach. FCCM 2015* 171–178. doi:10.1109/FCCM.2015.52

Smith S. 2009. QGModel. http://www.cims.nyu.edu/~shafer/tools/index.html (Accessed February 28, 2017).

Thornes T, Düben P, Palmer T. 2017. On the Use of Scale-Dependent Precision in Earth System Modelling. *Q. J. R. Meteorol. Soc.* doi:10.1002/qj.2974

Trader T. 2016. NVIDIA Unleashes Monster Pascal GPU Card at GTC16. *HPC Wire* https://www.hpcwire.com/2016/04/05/nvidia-monster-pascal-gpu-card-gtc16/ (Accessed April 11, 2016).

Tulloch R, Smith KS. 2006. A theory for the atmospheric energy spectrum: Depth-limited temperature anomalies at the tropopause. *PNAS* **103** : 14690–14694. doi:10.1073/pnas.0605494103

Vana F, Düben PD, Lang S, Palmer T, Leutbecher M, Salmond D, Carver G. 2017. Single Precision in Weather Forecasting Models: An Evaluation with the IFS. *Mon. Weather Rev.* **195** : 495–502. doi:10.1175/MWR-D-16-0228.1

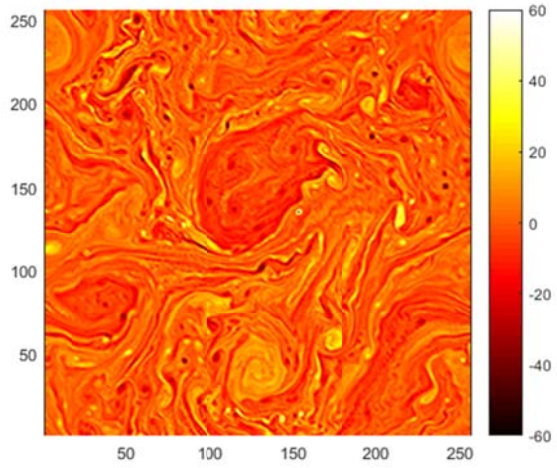

Figure 1: Example snapshot of the vorticity field under the short-term forecast setup. Here, the model has run for 80000 timesteps from the initial flat field. The flow passes a mountain of nondimensional height $H = 3$ in the centre of the domain.

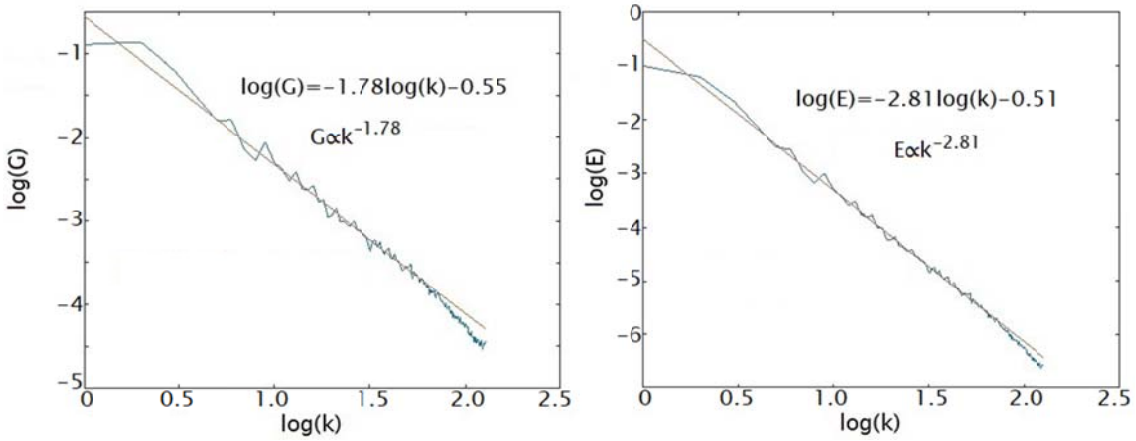

Figure 2: Enstrophy (left) and energy (right) spectra for our short-term forecasting setup, with best-fit straight lines, the equations for which are also shown. Both are base-10 logarithmic plots.

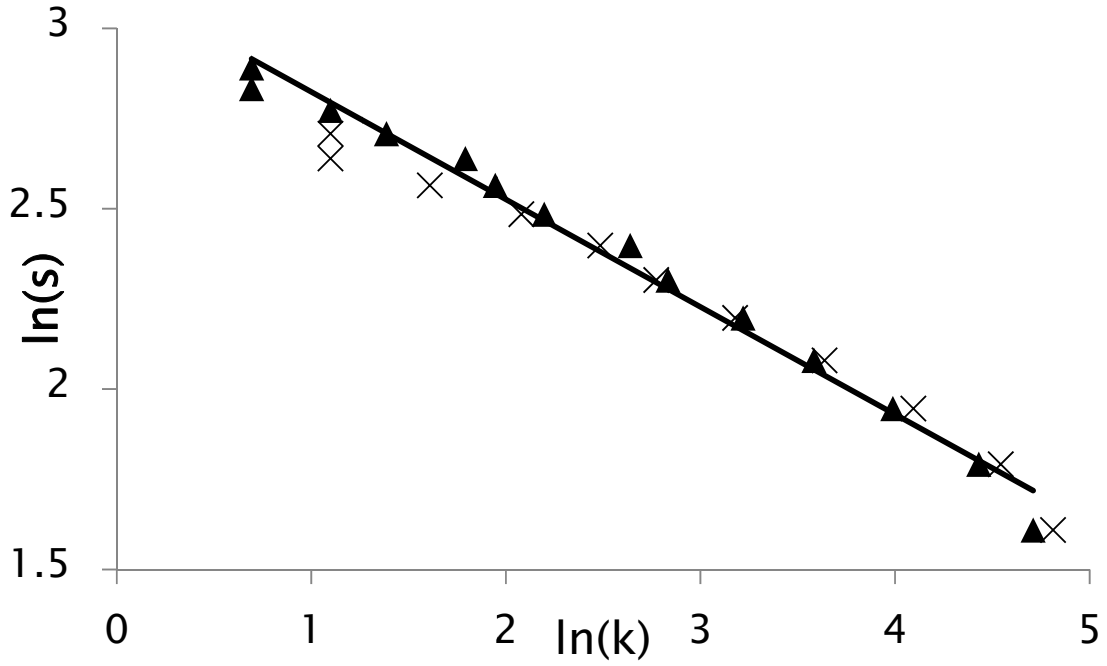

Figure 3: The natural logarithm of the number of bits in the significand *s* is plotted against that of the lowest wavenumber *k* that can be represented with that number of significand bits for the streamfunction (triangles) or vorticity (crosses) with negligible error (see Table 1). A best-fit straight line to the streamfunction points is plotted in black; a similar line for the vorticity is not plotted to preserve the clarity of the figure.

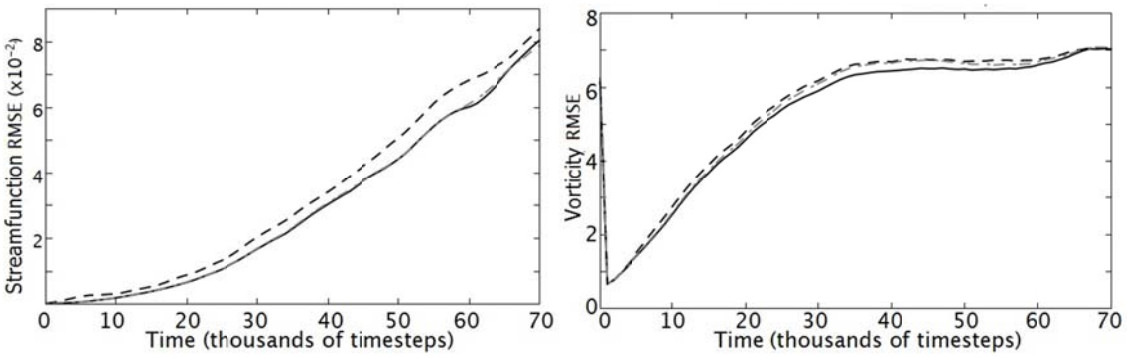

Figure 4: The RMSE of the 25-member ensemble mean in the streamfunction (left) and vorticity (right) is plotted for a double-precision control model (solid line), a scale-selective model optimised according to the vorticity column in Table 1 (grey dashed line), and a model with all wavenumbers reduced in precision to 15 bits in the significand (black dashed line) averaged over five runs of 70000 timesteps (100 days) each. The initial drop in error for the vorticity variable is due to model readjustment in response to the initial condition perturbations, which were introduced in the streamfunction only.

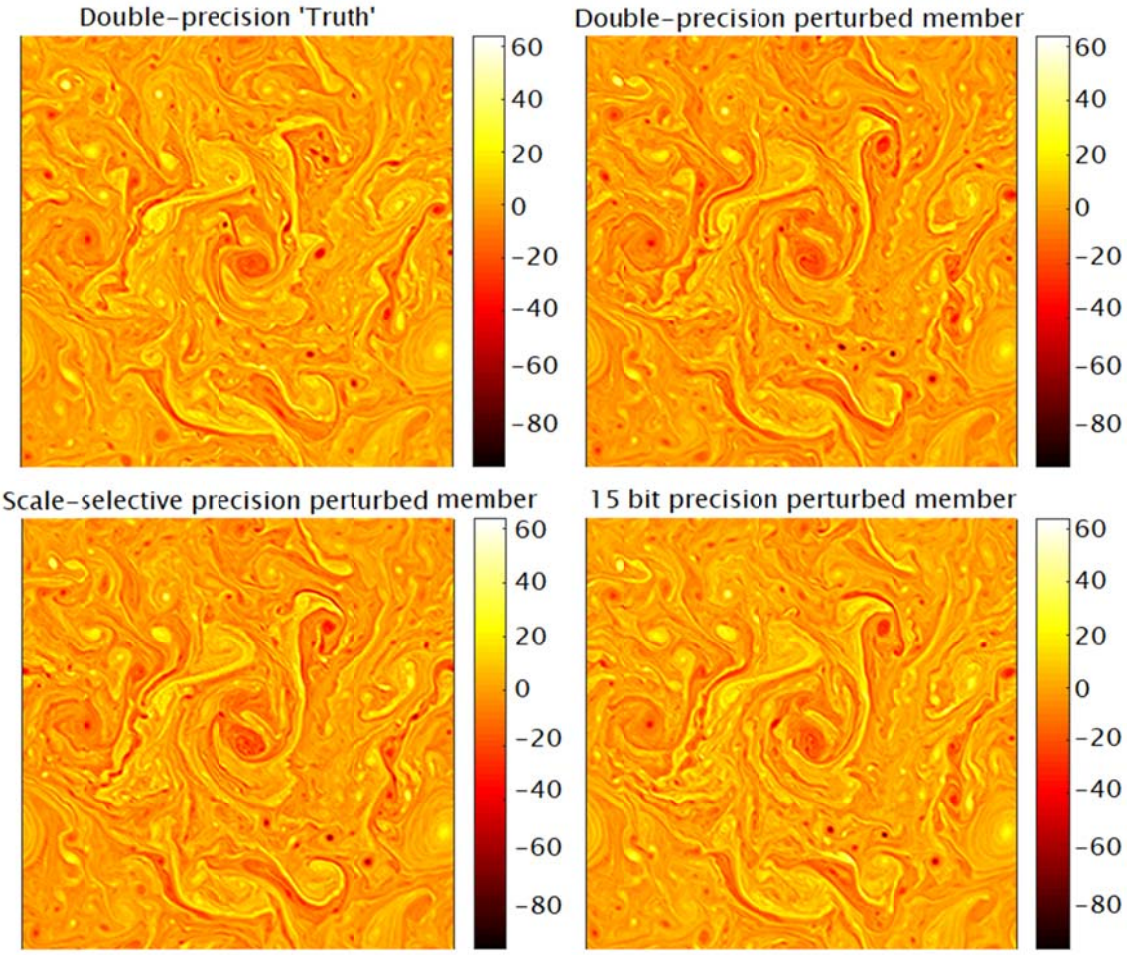

Figure 5: The vorticity field is plotted at 30000 timesteps after spin-up for the double-precision 'truth' (top left), and the first perturbed ensemble member for the double-precision control (top right), scale-selective (bottom left) and 15 bit precision (bottom right) models.

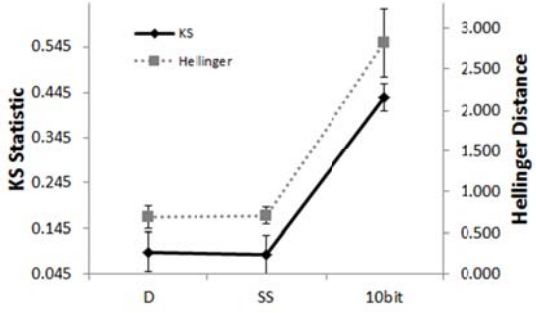

Figure 6: Kolmogorov-Smirnov (KS, solid line) and Hellinger Distance (dotted line) statistics are shown for a 250000 timestep run of the Double-precision (D), scale-selective (SS) and 10 bit models relative to the 'truth'. The values and error bars shown are mean values and standard errors computed from four independent runs.

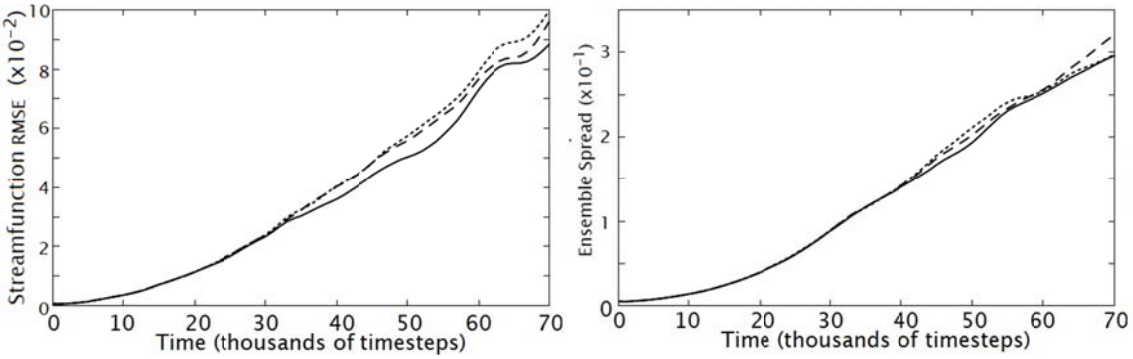

Figure 7: The Root Mean Square Error (RMSE, left) in the 10-member ensemble mean and the ensemble spread (right) are plotted for the streamfunction for a double-precision control model (solid line), and two models with the spectral part in scale-selective precision: the first with grid-point calculations carried out using 19 bits in the significand and FFTs with 11 bits in the significand (dotted line) and the second with FFTs and grid-point calculations in double-precision (dashed line). Simulations were carried out for 70000 timesteps after the initial spin-up to a single initial condition, and for each model ten perturbed ensemble members were used.

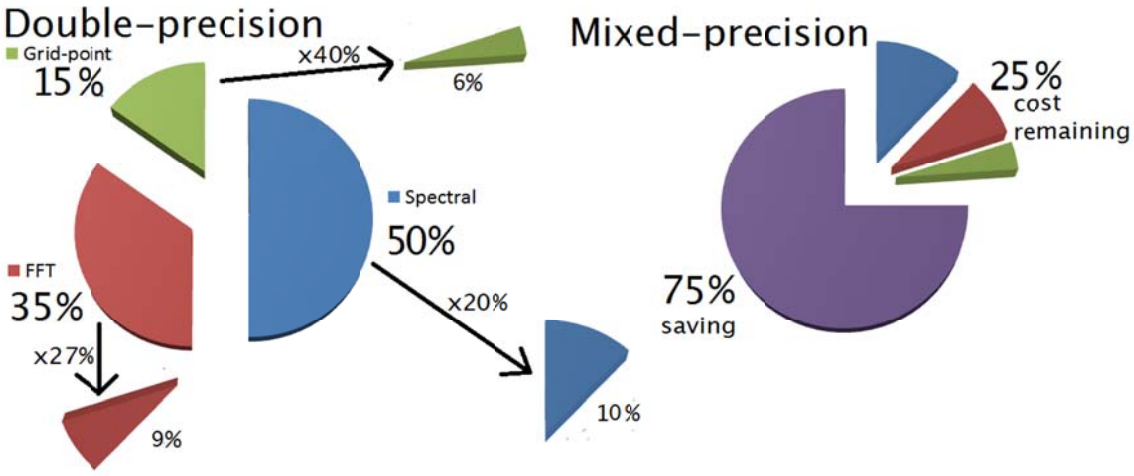

Figure 8: Diagram representing how the computational cost associated with the three components (spectral, grid-point and FFT) can be reduced using scale-selective significand precision for the spectral part, 19 bits in the significand for grid-point calculations and 11 significand bits for the FFT component, assuming no overhead costs.

Significand bits	Minimum k_t for accurate ψ	Minimum k_t for accurate ζ
1	NONE	NONE
2	NONE	NONE
3	NONE	NONE
4	NONE	NONE
5	111	123
6	84	94
7	54	60
8	36	38
9	25	24
10	17	16
11	14	12
12	9	8
13	7	5
14	6	3
15	4	3
16	3	0
17	2	0
18	2	0
19	0	0

Table 1: The minimum wavenumber of truncation k_t above which all wavenumbers in the streamfunction ψ and vorticity ζ can be represented with the specified number of bits in the significand with negligible difference in Absolute Error from a double-precision control, measured at 3000 timesteps under the short-term forecast setup.

s	c_i	$\psi\ n_i$	Cost %	$\zeta\ n_i$	Cost %
19	0.391	3	0.01	0	0
18	0.375	0	0	0	0
17	0.359	3	0.01	0	0
16	0.344	4	0.02	6	0.02
15	0.328	11	0.04	0	0
14	0.313	7	0.03	9	0.03
13	0.297	17	0.06	21	0.08
12	0.281	60	0.20	42	0.14
11	0.266	48	0.15	58	0.19
10	0.250	172	0.52	164	0.50
9	0.234	341	0.97	441	1.25
8	0.219	819	2.17	1089	2.89
7	0.203	2085	5.13	2635	6.48
6	0.188	2646	6.01	3161	7.18
5	0.172	2040	4.25	630	1.31
Sum		**8256**	**19.58**	**8256**	**20.07**

Table 2: For each number of bits in the significand s, the fractional cost c_i relative to a 64-bit control is computed, assuming 5 bits in the exponent and 1 bit for the sign of the reduced-precision number. The sum of all wavenumbers in two-dimensional wavenumber space, n_i, represented at each level of precision (see Table 1) is listed for cases where either the error in the streamfunction ψ or that in the vorticity ζ is used to determine at what range of wavenumbers each level of precision can be applied. $n_i/k_{\max}^2$ is multiplied by c_i to obtain the percentage cost associated with each band relative to the total cost across all wavenumbers in double-precision; the total cost is the sum of this over all bands, shown at the bottom of the table.